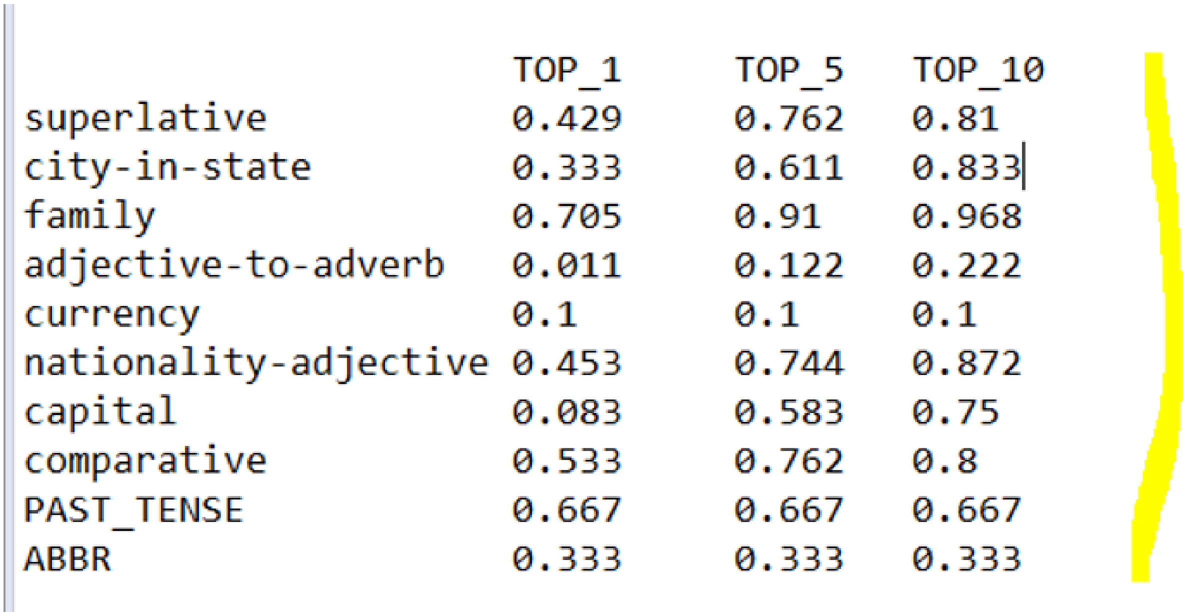
Report by: Anand Prashar

NLP Homework 5 – Extras

Q:

Try to improve your results on the analogy tasks in word-test.v3.txt and on the tasks you created in Q8.
You can do this by training your own vectors, downloading di_erent pretrained vectors, using a di_erent similarity metric, or using an entirely di_erent approach to the task (e.g. use WordNet). If you want to use pretrained vectors, some good families are (the following are hyperlinked):
_ Word2vec
_ GloVe
_ Dependency-based embeddings
Report on what you tried and what the e_ects were. Points will be awarded for creativity and insight.

Here is my accuracy result prior to making changes:



```
                        TOP_1     TOP_5    TOP_10
superlative             0.429     0.762    0.81
city-in-state           0.333     0.611    0.833
family                  0.705     0.91     0.968
adjective-to-adverb     0.011     0.122    0.222
currency                0.1       0.1      0.1
nationality-adjective   0.453     0.744    0.872
capital                 0.083     0.583    0.75
comparative             0.533     0.762    0.8
PAST_TENSE              0.667     0.667    0.667
ABBR                    0.333     0.333    0.333
```

I decided to use Word2Vec.

Downloaded a bin file ~1.5 GB from google word2Vec.

Instead of using pretrained vectors provided, I decided to build from original vectors of those words.

First I looked up words in file - word-test.v3.txt to know what words are needed.

And used set of them to build lookup objects from from gensim model.

This is how I got similarity:

```
most_similar = word_2_vec_gensim.similar_by_vector(w_vec, topn=15)
```

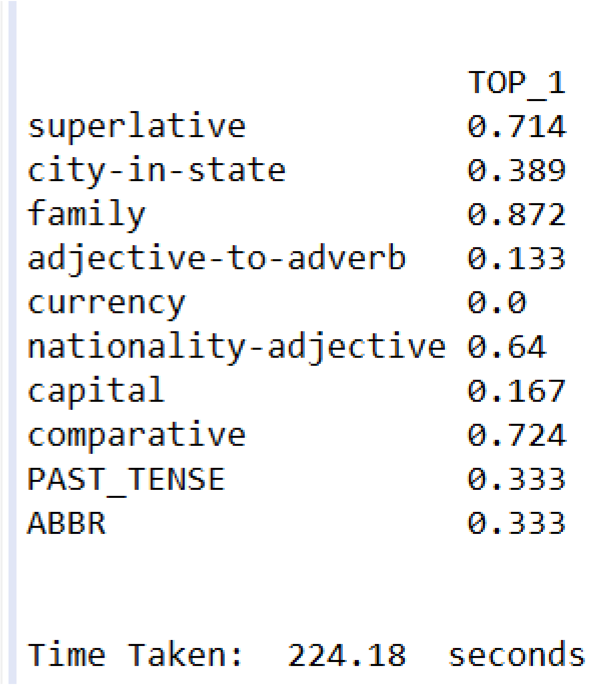
I then clipped to remove the set of words that should not be returned

I decided to increase the dimensionality from default from 100 to 300 for all words.

I also added stemmed version of some words to lookup to hit in case of not finding.

On running, I found very satisfactory results most likely due to wide range of contexts availability now

This is final result :



```
                      TOP_1      TOP_5    TOP_10
superlative           0.714      0.929    0.976
city-in-state         0.389      0.444    0.667
family                0.872      0.962    1.0
adjective-to-adverb   0.133      0.267    0.411
currency              0.0        0.0      0.0
nationality-adjective 0.64       0.767    0.779
capital               0.167      0.25     0.333
comparative           0.724      0.838    0.881
PAST_TENSE            0.333      0.667    0.667
ABBR                  0.333      0.667    0.667


Time Taken:  224.18  seconds
```

Except for the currency, there has been significant increase in accuracy in all sections !